SEMINARIO GRUPPO TEMATICO METODI e TECNICHE
La valutazione degli incentivi industriali: aspetti metodologici
Gruppo Tematico Metodi e Tecniche
Associazione Italiana Valutazione
Università di Brescia, 17 gennaio 2012
centro di studi e di ricerca dati metodi e sistemi

# Tecniche di Data Mining

**Marika Vezzoli**

Università di Brescia
Centro di Studi e Ricerca Dati Metodi Sistemi

# Introduzione

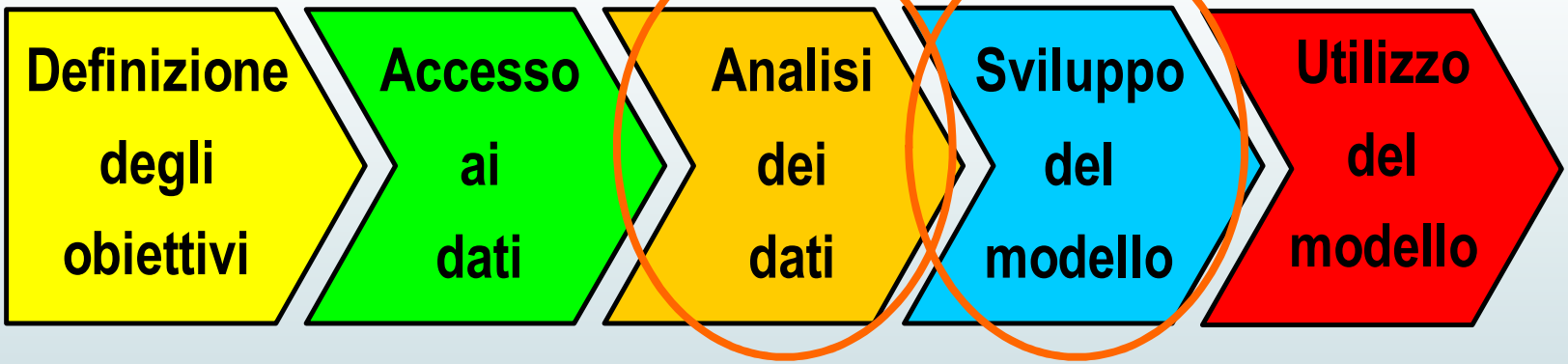
Il Data Mining è un processo dedicato alla

1. **selezione**
2. **esplorazione**
3. **analisi**

di database complessi e di dimensioni elevate con la finalità di ottenere informazioni utili nell'ambito di un generico problema decisionale

AV

# Introduzione




E' utile distinguere tra **obiettivi di tipo descrittivo** (riferiti alla situazione passata e/o attuale) e **obiettivi di tipo previsivo** (riferiti alla situazione futura)

E' inoltre fondamentale tenere presente la distinzione tra **variabili dipendenti o obiettivo** ($Y$) e **variabili indipendenti o predittori** del problema oggetto di studio

# Introduzione

Variabile dipendente qualitativa →

**Modelli di CLASSIFICAZIONE**

Variabile dipendete quantitativa →

**Modelli di REGRESSIONE**

# Introduzione

| Modelli di Regressione | Modelli di Classificazione |
|---|---|
| Regressione lineare/non lineare | Regressione logistica |
| **CART** | |
| Reti neurali di regressione | Reti neurali di classificazione |
| | Modelli Naive Bayes |

# Quando si utilizza il CART?

1. **Analizzare relazioni complesse, non lineari** e difficili da identificare
2. **Selezionare le variabili più importanti** nella spiegazione di fenomeni «misurabili»
3. **Prevedere** in maniera semplice ed intuiva **senza imporre** particolari **ipotesi sulla forma distributiva** delle variabili utilizzate nell'analisi

**Ambiti di applicazione → molteplici** (economico-finanziario, medico, fisico, ambientale, meteorologico, etc.)

# Vantaggi

1. Algoritmo relativamente **veloce** (in termini computazionali)
2. In grado di trattare **tutti i tipi di variabili**
3. Invariante alle **trasformazioni** monotone
4. **Interpretabilità** (se l'albero è … "bonsai")
5. Capacità di gestire dati con problematiche serie:
   - ✓ ***Missing values***
   - ✓ **Correlazione tra le variabili**

# Svantaggi

1. L'algoritmo è di tipo **sequenziale** (ricorsivo)
   → le scelte effettuate ad un passo influenzano anche quelle nei passi successivi
2. I risultati sono **instabili**
3. Il guadagno in termini di **accuratezza** è **modesto** rispetto alla regressione logistica

AV

# CART

- È una tecnica non parametrica, che permette la segmentazione delle osservazioni utilizzando strutture grafiche gerarchiche chiamate appunto “alberi”
- Sviluppata da Breiman Friedman Olshen e Stone (1984)
- Seleziona le variabili più importanti nel determinare la variabile di risposta $Y$ (***variable selection***)

# CART

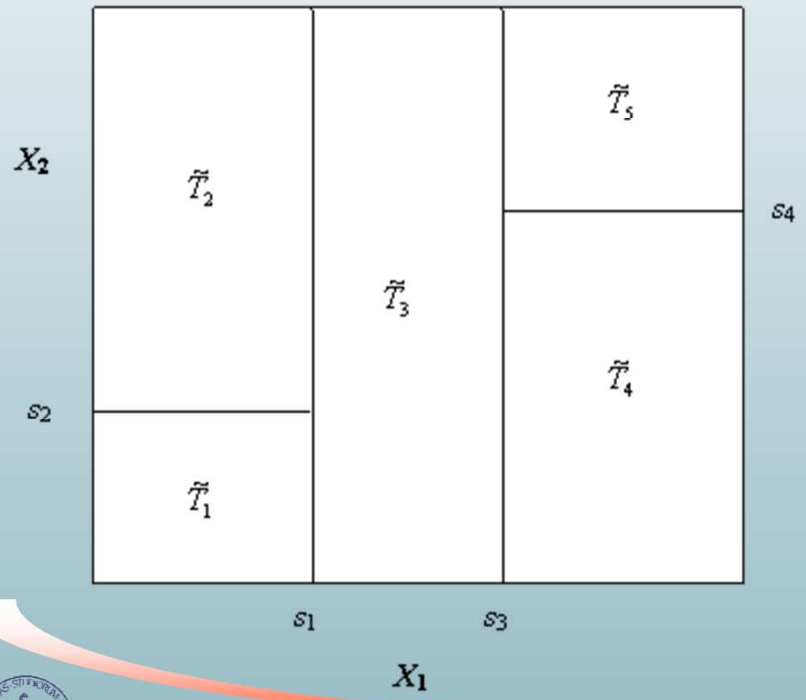
- Il CART è un algoritmo che partiziona lo spazio $S$ delle variabili input $X$ in una serie di regioni omogenee e disgiunte rispetto alla variabile obiettivo $Y$ stimando un modello semplice in ognuna di esse

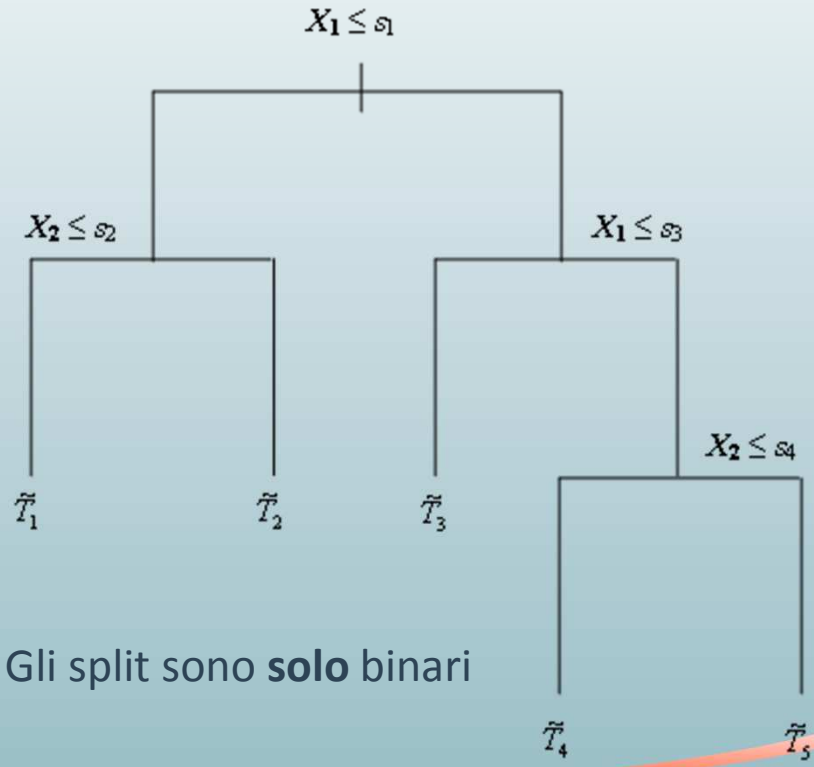
  a) **alberi di regressione**: prevede $Y$ con una costante $c_m$ nella regione

  b) **alberi di classificazione**: attribuiscono una classe ad ogni regione (***majority rule***)

# Alberi di Regressione

Supponiamo di avere una variabile risposta $Y$ e 2 soli regressori (continui). In questo modo possiamo rappresentare graficamente lo spazio $S$:




Gli split sono **solo** binari

# Alberi di Regressione

L'algoritmo identifica automaticamente le variabili e i punti di *split*. Ogni regione con $m = 1,\ldots, M$ è rappresentata da una costante $c_m$

$$\hat{c}_m = \bar{y}_i = \frac{1}{N_m} \sum_{\mathbf{x}_i \in \tilde{T}_m} y_i$$

dove $N_m$ è il numero di osservazioni nel nodo $m$

La costante corrisponde pertanto alla media degli $y_i$ appartenenti al nodo

# Alberi di Regressione

Per misurare la variabilità nel nodo $m$ si utilizza la seguente funzione di perdita

$$R(m) = \frac{1}{N_m} \sum_{\mathbf{x}_i \in \tilde{T}_m} (y_i - \hat{c}_m)^2$$

Quando il nodo $m$ viene diviso in due nodi figli ($m_1$ e $m_2$), il **miglior *split*** $s^*$ è quello che **massimizza la riduzione dell'errore $R(m)$**

# Alberi di Classificazione

Negli alberi di classificazione la variabile di risposta $Y$ è categorica ($Y = \{1,2,\ldots,k,\ldots K\}$)

Lo scopo è di modellare la probabilità di appartenere ad ognuna delle $k$ categorie della variabile di risposta $Y$ condizionatamente ai predittori $\mathbf{X} = (X_1,\ldots,X_{\mathcal{R}})$

A tal fine si ricorre alla proporzione di osservazioni di classe $k$ nel nodo $m$

$$\hat{p}(k \mid m) = \frac{1}{N_m} \sum_{\mathbf{x}_i \in \tilde{T}_m} I(y_i = k)$$

Ad ogni nodo è assegnata la classe $k$ maggiormente rappresentativa (*majority rule*)

# Alberi di Classificazione

**Esempio:** in una foglia *m* cadono 10 osservazioni così ripartite fra le 4 categorie della variabile dipendente:

- 0 → categoria 1
- 1 → categoria 2
- 6 → categoria 3
- 3 → categoria 4

Quindi:

$$\hat{p}(1 \mid m) = \frac{1}{10} \times 0 = 0 \qquad \hat{p}(2 \mid m) = \frac{1}{10} \times 1 = 0.1 \qquad \hat{p}(3 \mid m) = \frac{1}{10} \times 6 = 0.6 \qquad \hat{p}(4 \mid m) = \frac{1}{10} \times 3 = 0.3$$

e al nodo *m* si attribuisce la categoria 3 (*majority rule*)

$$k(m) = \max_k \hat{p}(k \mid m) = 3$$

# Alberi di Classificazione

Analogamente agli alberi di regressione, si sceglie lo *split* che ad ogni nodo minimizza il ***tasso di errata classificazione***, ovvero la probabilità di classificare con $k$ una osservazione di classe $k'$. Ad esempio, nel nodo $m$:

$$R(m) = \frac{1}{N_m} \sum_{\mathbf{x}_i \in \tilde{T}_m} I\left(y_i \neq k(m)\right) = 1 - \hat{p}\left(k \mid m\right)$$

Breiman Friedman Olshen e Stone (1984) suggerirono altre misure di impurità poiché dimostrarono che il tasso di errata classificazione non genera alberi ottimali. Alternative largamente utilizzate sono ***l'indice di eterogeneità di Gini*** o ***l'entropia***

# Scegliere la grandezza dell'albero

Un tipico problema degli alberi: l'***overfitting***

Esistono varie soluzioni: imporre la grandezza minima dei nodi finali, imporre il numero di nodi finali ecc.

Breiman e Stone (1978) proposero il metodo del ***pruning***. Questo metodo genera una sequenza ottimale (nidificata) di sottoalberi potati tra i quali si seleziona l'albero in grado di tener conto sia del costo legato alla potatura sia del beneficio ottenuto nell'interpretazione di un albero di piccole dimensioni

# Se $Y$ dummy → Alberi di Regressione

Quando la variabile indipendente è una *dummy* ($Y = \{0,1\}$) è possibile usare indistintamente gli alberi di regressione o classificazione. Infatti la media degli $y_i$ nel nodo (ovvero la previsione ottenuta utilizzando un albero di regressione) coincide con la proporzione di osservazioni di classe $1$ nel nodo preso in considerazione

Questa caratteristica è interessante perché in certe situazioni gli alberi di regressione risultano più informativi rispetto a quelli di classificazione

AV

# Esempio: il rischio sovrano

Il *dataset* è stato fornito dall'**I***nternational* **M***onetary* **F***und* (**IMF**) ed è relativo a 66 paesi emergenti, rilevati su un periodo dal 1975 al 2002 (Manasse and Roubini, 2005)

I paesi sono classificati in ***default*** ad una certa data quando non sono in grado di onorare il debito emesso (Standard&Poor's)

**Condizione posta sui dati**: l'anno precedente il paese non doveva essere in *default* (si è quindi generato un campione *non bilanciato,* ovvero ogni unità nel panel data ha un numero diverso di osservazioni)

# Esempio: il rischio sovrano

Le variabili esplicative utilizzate nell'esempio:

- IMF: prestito concesso dall'IMF (*dummy*)
- CAY: bilancia corrente
- ResG: crescita delle riserve
- WX: esportazioni
- WXG: crescita delle esportazioni
- STDR: debito a breve termine sulle riserve
- M2R: rapporto tra M2 e riserve
- OVER: dev.std del tasso di cambio
- EXCHR: tasso di cambio

# Esempio: il rischio sovrano

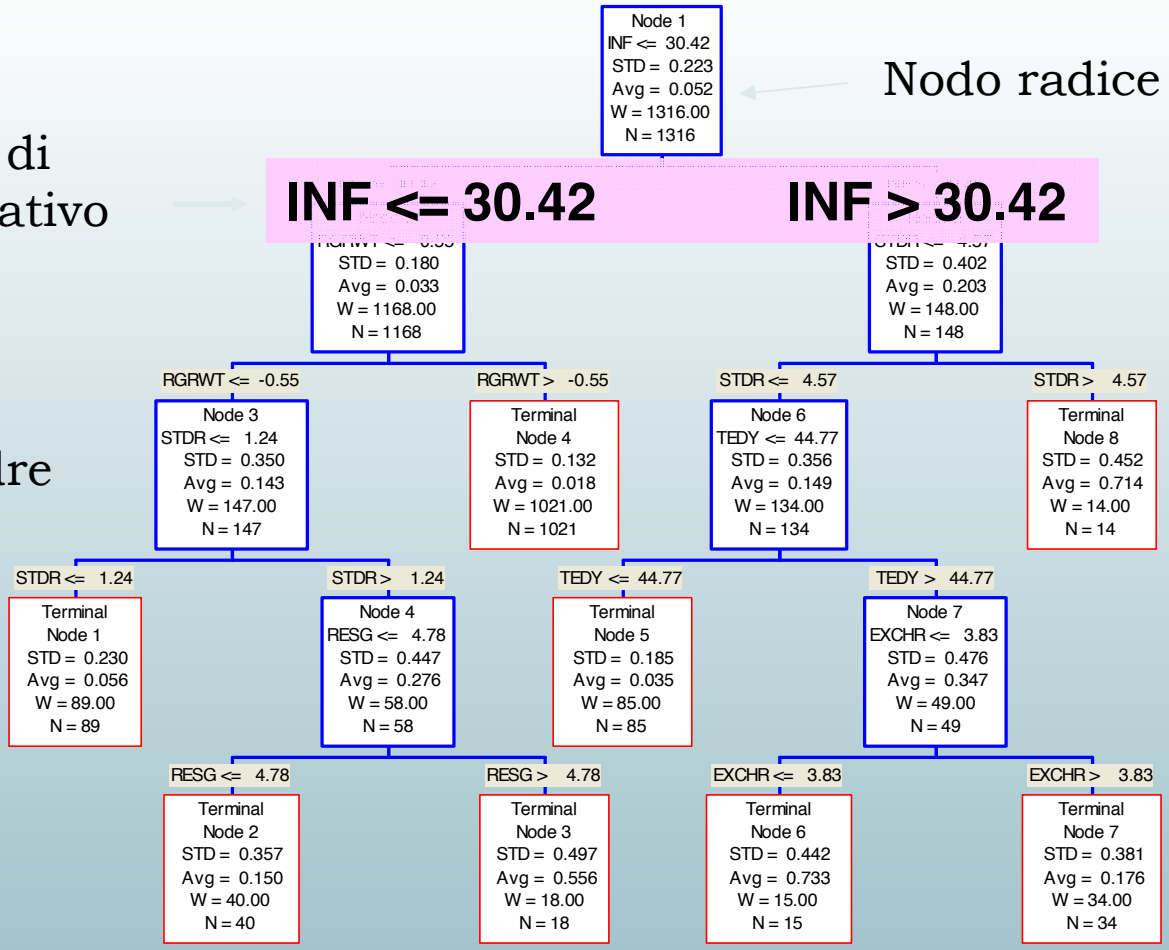
- INF: inflazione
- NRGWT: crescita nominale PIL
- RGRWT: crescita reale PIL
- TEDY: debito totale esterno
- PR: diritti politici (variabile categoriale) -Class. IMF
- MAC: apertura al mercato (*dummy*)
- DAFR, DADV, DAPD, DWHD, DMED, DTRANS: variabili regionali (*dummy*)
- DOIL: petrolio (*dummy*)

# Esempio: il rischio sovrano

Nodo radice

Variabile di *split* e relativo valore *s**

**BLU**: nodi padre




**ROSSO**: nodi finali (foglie)

# Una soluzione all'instabilità

Tra gli svantaggi del CART vi è l'**instabilità** dei risultati

Recentemente sono state proposte delle tecniche, denominate ***Ensemble Learning***, che cercano di fornire una soluzione a questa problematica

L'idea di base è di perturbare ripetutamente il dataset di partenza in modo da crescere un numero elevato di alberi, combinando infine i risultati ottenuti

È stato dimostrato che **le previsioni ottenute sono più accurate e stabili** di quelle ottenute crescendo un solo albero

Le tecniche di *ensemble learning* maggiormente conosciute sono:

- ✓ **Bagging (Breiman, 1996)**
- ✓ **Boostig (Freund and Schapire, 1996)**
- ✓ **Random Forest (Breiman, 2001)**

# Misure di importanza

Con l'introduzione degli *ensemble learning* sono state sviluppate le cosiddette **misure di importanza** (**Variable Importance Measures**)

Tramite queste misure si ottiene un ***ranking*** (ordinamento) delle variabili → dalla più importante a quella meno importante

In presenza di dataset con un numero elevato di variabili è interessante utilizzare tali tecniche per circoscrivere il sottoinsieme di variabili maggiormente esplicative del fenomeno esaminato

# Conclusioni

Le tecniche di *data mining* **processano "in modo democratico "** dati di complessità anche elevata superando limitazioni connesse agli approcci tradizionali, spesso concepiti nell'illusione di un mondo Gaussiano e lineare

La democraticità con cui sono processati i dati si traduce in una **impostazione di analisi che prescinde da qualunque teoria a priori**, senza dunque soffrire di assunzioni che spesso producono risultati tautologici, validi per descrivere il passato ma fallaci nel formulare previsioni future

E' tuttavia **indispensabile possedere una expertise elevata** per poter correttamente calibrare tali tecniche alle singole problematiche da affrontare